Künstliche Intelligenz und menschlicher Verstand: Grundprobleme psychologisch orientierter KI-Forschung

Manhart, Klaus

Veröffentlichungsversion / Published Version
Zeitschriftenartikel / journal article

Psychologische Beiträge, Band 33, 1991, S. 281-313

Künstliche Intelligenz und menschlicher Verstand
Grundprobleme psychologisch orientierter KI-Forschung[1]

K. MANHART[2]

Zusammenfassung, Summary

Verschiedene Computerwissenschaftler und Psychologen glauben, daß mit der Künstlichen Intelligenz (KI) die alten Rätsel um den menschlichen Verstand gelöst werden können. Den Grund sehen viele darin, daß Computer als Symbolverarbeitungsmaschinen beeindruckende Gemeinsamkeiten mit dem menschlichen Geist aufweisen und sich in beiden Systemen dieselben theoretischen Strukturen realisieren. Vor diesem Hintergrund beschwören Philosophen die real existierende Denkmaschine herauf, und manche Psychologen glauben, mit Hilfe der KI Einblicke in die Funktionsweise des menschlichen Verstandes zu bekommen. Probleme beider Disziplinen sind damit eng miteinander verwoben. Der Artikel versucht, die wichtigsten Probleme der heutigen kognitionswissenschaftlichen KI im philosophisch-psychologischen Kontext zusammenfassend darzustellen.

Artificial Intelligence and the human mind
Basic problems of psychologically oriented AI-research

Some computer scientists and psychologists believe that the old puzzles concerning the human mind could be solved with Artificial Intelligence (AI). For most of them, the reason is that computers as symbolprocessing machines have impressive common grounds with the human mind and the same theoretical structures are realised in both systems. Against this background, philosophers evoke the real existing thinking machine and psychologists believe to discover insights into the human mind with the help of AI. Thus, problems of both disciplines are linked close together. The intent of this paper is to summarize the most important problems of modern cognitive AI in the context of philosophy and psychology.

[1]Die Herausgeber der Psychologischen Beiträge möchten deren Leserkreis mit dem Abdruck dieses Artikels auf die Entwicklungen in diesem für die Psychologie zunehmend wichtigen Bereich aufmerksam machen.

[2]Dipl.Soz. Klaus Manhart, Institut für Soziologie der Ludwig-Maximilians-Universität München, Konradstr. 6, W-8000 München 40

人工知能と人間の知性：心理学的ＡＩ研究の基礎的問題

さまざまなコンピュータ科学者や心理学者は，人工知能（ＡＩ）によって人間の知性に関わる古い謎を解くことができると信じている．記号処理機械としてのコンピュータは人間の精神と深い共通性を持ち，両システムには同じ理論的構造が実現される，というのがその理由である．この背景のために，哲学者は真に存在する思考機械を思い起こし，心理学者はＡＩの助けを借りて人間の知性を解き明かすことができると信じている．つまり，両者の問題は密接に関連している．本論文では，現在の認知科学的ＡＩの最も重要な問題を哲学と心理学の文脈でまとめることを試みる．

（山下利之　Dr. Toshiyuki Yamashita）

1. Einleitung

Die Frage, was Denken ist und wie der Verstand funktioniert, beschäftigte Philosophen seit der Antike - ohne eine Antwort gefunden zu haben. Im 20. Jahrhundert machten sich empirische Wissenschaften wie Psychologie oder Neurologie an diese Aufgabe heran und waren bislang kaum erfolgreicher. Über 2000 Jahre philosophischen Studiums und 100 Jahre Psychologie/Neurologie reichten nicht aus, um die Mechanismen der menschlichen Intelligenz zu ergründen. Wir verstehen die Struktur des Universums und den Mikrokosmos der Atome, aber Fragen nach dem uns nächstliegenden - nach der Natur des menschlichen Denkens - können im 20.Jahrhundert immer noch nicht beantwortet werden.

Nach Meinung einiger bedeutender KI-Theoretiker und Philosophen stehen wir nun aber kurz davor, elementare Einsichten in die menschliche Intelligenz gewinnen zu können. Der Grund ist die KI. Der KI-Philosoph John Haugeland deutet in der Einleitung zu seinem Buch "Künstliche Intelligenz - Programmierte Vernunft" theatralisch die Rolle der KI bei diesem Unternehmen an:

"Was ist Verstand? Was ist Denken? Was ist es, das dem Menschen seine Sonderstellung im ganzen uns bekannten Universum verleiht? Von Fragen wie diesen werden die Philosophen seit Jahrtausenden gequält; ihre Fortschritte waren jedoch (jedenfalls nach wissenschaftlichen Maßstäben) eher gering - bis vor kurzem. Denn die heutige Generation hat ein plötzliches und brillantes Aufblühen der Philosophie und der Wissenschaft des Verstandes erlebt; inzwischen befinden sich nicht nur die Psychologie, sondern auch eine Fülle verwandter Disziplinen in den Geburtswehen einer großartigen geistigen Revolution. Der Inbegriff dieser dramatischen Entwicklung ist die *Künstliche Intelligenz*, jene aufregende und neuartige Anstrengung, Computern das Denken beizubringen. Das grundlegende Ziel dieser Forschung ist nicht etwa nur, Intelligenz zu simulieren oder irgendeine raffinierte Imitation hervorzubringen. Nein, "KI" will die Sache selbst: *Maschinen mit Verstand*, im vollen und wörtlichen Sinne. Das ist keine Science-fiction, sondern reale Wissenschaft. Sie beruht auf einem theoretischen Konzept, das ebenso tiefgreifend wie gewagt ist: Im Grunde genommen sind wir *selbst Computer*. Dieser Gedanke - der Gedanke, daß Denken und die Tätigkeit eines Computers im Prinzip

dasselbe sind - ist das Thema dieses Buches" (Haugeland 1987:2, Hervorhebungen von Haugeland).

Haugeland spricht in dieser Einleitung zentrale Thesen der philosophischen KI an:
-die KI ist ein neues, revolutionäres Paradigma in den Humanwissenschaften;
-KI will Computern das Denken beibringen und Maschinen mit Verstand bauen;
-Denken und das, was Computer tun, ist im Grunde dasselbe.

Wichtig ist, daß Haugeland seine Behauptungen nicht metaphorisch meint, sondern wörtlich. KI-Computer sollen - wenn nicht heute, aber irgendwann - einen Verstand haben und denken wie Sie und ich.

Seine Thesen gründen auf der Annahme, daß wir mit dem modernen Digitalcomputer ein Werkzeug in der Hand haben, welches auf der gleichen Basis funktioniert wie das menschliche Gehirn: Computer manipulieren Symbole. Sie tun damit etwas, was große Ähnlichkeiten mit dem menschlichen Geist aufweist. Denken ist nämlich nichts anderes als die "rationale Manipulation von Symbolen", eine Auffassung, die tief in der abendländischen philosophischen Tradition verwurzelt ist und von Thomas Hobbes (1650) formuliert wurde.

Der englische Philosoph Thomas Hobbes (1588-1679) betrachtet Denken als "geistigen Diskurs", als "symbolische Operationen", die wie lautes Diskutieren oder Rechnen mit Papier und Bleistift geschehen, nur innerlich. *Rationales* Denken folgt darüberhinaus methodischen Regeln, so wie sich etwa ein Buchhalter an Regeln der numerischen Mathematik hält. Denken funktioniert demzufolge wie ein geistiger Abakus, bei dem kleine Teile streng nach Regeln des Verstandes hin- und hergeschoben werden. Diese Auffassung ist eine zentrale philosophische Grundlage der heutigen KI. (Haugeland 1987: 19-20).

Die Tatsache, daß Computer und Menschen im Grunde genommen dasselbe tun - nämlich Symbole zu verarbeiten - sehen bestimmte Forscher als notwendige und hinreichende Voraussetzung dafür, Maschinen zu bauen, die zu ähnlichen kognitiven Leistungen fähig sind wie der Mensch. Wenn wir in der Lage sind, solche Maschinen zu bauen, können wir damit aber auch mehr über den menschlichen Geist erfahren. Mit den Programmen, die dies ermöglichen, können nämlich dann Rückschlüsse auf die Funktionsweise des menschlichen Geistes gemacht werden. Denn wir verstehen zwar den Geist nicht, aber wir verstehen Computer (-programme) und wenn wir Computerprogramme verstehen, die den Geist nachbilden, so verstehen wir auch den Geist. Die KI-Maschine ermöglicht es somit quasi, 2 Fliegen mit einem Streich zu erledigen: Maschinen mit menschenähnlicher Intelligenz zu erschaffen und den menschlichen Geist besser zu verstehen.

Wissenschaftler, die solche Ziele verfolgen, werden zweifellos der KI zugerechnet. Um einem Mißverständnis vorzubeugen: die meisten KI-Forscher werden sich von

diesen Motiven distanzieren (falls sie überhaupt davon etwas wissen); sie wollen lediglich nützliche, praktisch einsetzbare Programme schreiben, ohne Erkenntnisse über den menschlichen Geist gewinnen zu wollen oder die Super-Denkmaschine zu kreieren.

Die heutige KI-Gemeinde läßt sich in 2 Lager aufteilen, deren Mitglieder in ihren Zielsetzungen nicht viel gemeinsam haben. Grob gesprochen, verfolgt die eine Gruppe das Ziel, Computer einfach nützlicher zu machen, d.h. intelligente praxisrelevante Maschinen zu konstruieren, ohne sich darum zu kümmern, ob Menschen diese Aufgabe kognitiv genauso oder ähnlich ausführen. Das explizite Ziel ist also die Entwicklung praktisch anwendbarer, leistungsfähiger und möglichst effizienter Problemlösungssysteme, unabhängig davon, wie Menschen Probleme lösen. Ein typisches Beispiel sind Expertensysteme, bei denen es in der Regel nicht darauf ankommt, das kognitive Verhalten von Experten exakt nachzubilden. Diese Gruppe besteht hauptsächlich aus Computerwissenschaftlern, die KI als "harte" technologische Teildisziplin der Informatik betrachten.

Diesem "produktorientierten" Ansatz stehen theoretisch ausgerichtete KI-Forscher mit einem völlig anderen Motiv gegenüber. Erklärtes Ziel dieses Ansatzes ist es, menschliches Denken und Verhalten mit Hilfe von Computern und Programmen besser zu verstehen. Es werden Modelle von Denkprozessen auf Maschinen nachgebildet und aus diesen Modellen Rückschlüsse auf die Funktionsweise beim Menschen durchgeführt. Dieser Ansatz strebt also nicht danach, praktisch verwendbare, auf Leistung angelegte Programme zu erstellen, sondern Programme, die menschliches Denken und Verhalten möglichst genau modellieren.

Theoretisch orientierte KI-Forscher bilden eine relativ heterogene Gruppe aus unterschiedlichsten Disziplinen wie Psychologie, Linguistik, Informatik etc. Es besteht die Tendenz, die KI eher als ein Teilgebiet der Humanwissenschaften zu betrachten, wie etwa Waltz in seinem Überblicksartikel im "Scientific American" schreibt:

"Einige Forscher behaupten nichtsdestoweniger, die Wissenschaft von der Künstlichen Intelligenz sei, grob gesprochen, letztlich nur ein Zweig der Psychologie des Menschen. Sie gehen davon aus, daß jede erfolgreiche Simulation einer intellektuellen Leistung des Menschen durch ein Computerprogramm einen "Existenzbeweis" für das Computermodell der menschlichen Intelligenz erbringe. Auch wenn sich der Ablauf des Programms so sehr von mentalen Prozessen unterscheide, daß es keinen direkten Beitrag zur Psychologie zu leisten vermöge, könne es doch zumindest gewichtige Aspekte der menschlichen Intelligenz simulieren, etwa die Schwierigkeit, ihre Leistungen genau vorherzusagen" (Waltz 1982:70-71).

Ein extremes Beispiel für ein dem Theoriemodus zuzuordnendes Programm ist EPAM (Elemantary Perceiver and Memorizer) von Feigenbaum (1963), das das Silbenlernen von Menschen nachahmt.

Das Programm prägt sich Paare unsinniger Silben ein. Diese Aufgabe ist für einen Computer trivial: man speichert einfach die Reizsilben und die zugehörigen Antwortsilben ab und der Computer gibt bei Eingabe einer Reizsilbe die dazugehörige Antwortsilbe aus. Menschen aber sind vergeßlich und können sich - im Gegensatz zu Maschinen - nicht millionenweise solcher Paare merken. EPAM simuliert nun genau diese menschliche Eigenschaft. Ein solches Programm ist technologisch natürlich völlig wertlos, für Psychologen aber sehr interessant, als sie ihre Theorien damit überprüfen können.

Die meisten philosophisch angehauchten KI-Forscher - und dazu gehören so bedeutende KI-Pioniere wie Herbert Simon oder Marvin Minsky - stehen aber hinter den von Haugeland angesprochenen Zielen und Thesen, so verrückt sie auch scheinen mögen.

Wie in der Philosophie üblich, sind solche Thesen natürlich nicht allgemein akzeptiert und erscheinen für manche Leser auf den ersten Blick wahrscheinlich auch reichlich versponnen. Wir werden ihre Grundlagen im folgenden detaillierter betrachten, so daß sie vielleicht etwas verständlicher erscheinen. Die Theorie, die den Hintergrund hierzu liefert und versucht, eine Antwort zu geben auf die Frage nach dem menschlichen Geist, ist die sog. funktionalistische Geistestheorie. Bevor wir darauf eingehen, betrachten wir aber eine andere, sehr einflußreiche "Geistestheorie", die allerdings ganz ohne Geist auskommt.

2. Der Behavourismus

In den fünfziger und sechziger Jahren waren die Humanwissenschaften, insbesondere die Psychologie, weitgehend vom Paradigma des Behaviourismus beherrscht. Das behaviouristische Programm wurde Anfang des Jahrhunderts von J.B.Watson formuliert und von B.F. Skinner und seinen Anhängern ausgearbeitet. Nach der Grundthese des Behaviourismus kann das Reden über mentale (geistige) Zustände nicht nur vermieden werden, indem über Verhalten gesprochen wird, das Reden über mentale Zustände ist strikt genommen sogar sinnlos. Skinner ging davon aus, daß das einzig relevante Ziel in der Psychologie die Voraussage des Verhaltens ist. Für die Voraussage des Verhaltens ist es aber - so Skinner - nicht notwendig, sich auf mentale Phänomene zu beziehen. Es mag zwar mentale Phänomene geben, diese sind aber völlig irrelevant für die Entwicklung von Verhaltensgesetzen. Diese psychologischen Thesen wurden später von Ryle in seinem berühmten Buch "Der Begriff des Geistes" philosophisch untermauert. Ryle versuchte darin den Cartesianischen Dualismus - nach dem es sowohl physische als auch mentale Phänomene gibt - zu widerlegen und zu zeigen, daß lediglich das Reden über physische Ereignisse sinnvoll ist.

Nach Descartes (1596-1650) existieren in der Welt zwei völlig unterschiedliche Substanzen: Geist und Materie. Materie ist etwas physisches, das immer räumlich vorhanden ist und eine eindeutig bestimmbare Größe und Form hat, während der Geist etwas nicht physisches, eben "geistiges" ist, der mit der materiellen Welt nichts zu tun hat. Die Dualisten stehen vor einem Riesenproblem: wie können zwei Welten, die nichts miteinander zu tun haben, aufeinander einwirken. Denn wenn ich mich verletze (physische Welt), so hat dies den geistigen Zustand "Schmerzen" (psychische Welt) zur Folge und umgekehrt: wenn mein Geist entschließt, den Arm zu heben, so hat das entsprechende Auswirkungen auf die materielle Welt, indem mein Arm sich nach oben bewegt. Dieses "Geist-Körper-Problem" führte in der Philosophie zu verzweifelten Denkkapriolen.

Heyer (1988) erinnern die Thesen des Behaviourismus an Exorzismus, bei dem einem der so vertraute Geist mit beschwörenden Formeln ausgeredet werden soll; die wesentlichen Einwände gegen den Behaviourismus lauten (Heyer 1988a: 36-37, vgl. auch Bieri 1981: 31-32):

"1)Wir erkennen uns nicht wieder in dem Bild, das der Behaviourismus von unseren psychischen Vorgängen zeichnet; Erinnerungen, Gefühle oder Träume spielen de facto eine bedeutsame kausale Rolle für unser Verhalten und die Art und Weise, wie wir unser Verhalten begründen.

2)Die Beschreibung mentaler Zustände kann nicht punktuell nach dem Muster von Reiz und Reaktion geschehen; der Behaviourismus verkennt die holistische Natur des Geistes.

3)Rein empirisch gesprochen sind Theorien, die sich auf mentale Phänomene als Ursachen von Verhalten beziehen, fruchtbarer als Theorien, die behaviouristischen Beschränkungen unterliegen."

Der Behavourismus hat inzwischen an Bedeutung verloren und wurde weitgehend zurückgedrängt. Man hat eingesehen, daß mentale Zustände und Ereignisse in der Welt eine kausale Rolle spielen und sich nicht in Verhaltensbeschreibungen auflösen lassen. Daß der Behaviourismus zurückgedrängt wurde, ist nicht zuletzt ein Erfolg der KI und kognitiven Wissenschaften. Diese Disziplinen zeigen, daß es sinnvoll und in exakter Weise möglich ist, sich auf kognitive Prozesse zu beziehen ohne behaviouristischen Tendenzen zu verfallen (Heyer 1988a).

3. Die funktionalistische Geistestheorie

Der Funktionalismus als philosophische Theorie des Geistes wurde von dem Mathematiker und Philosophen Hilary Putnam (1960) in seinem Aufsatz "Minds and Machines" entwickelt (deutsche Übersetzung "Geist und Maschine" in Beckermann 1985a). Die funktionalistische Geistestheorie bringt sozusagen den vom Behaviourismus ausgetrie-

benen Geist wieder zurück in die Humanwissenschaften und stellt bestimmte Korrespondenzen her zwischen der Funktionsweise des menschlichen Geistes und der Funktionsweise von Computern.

Dem Funktionalismus unterliegt die These, daß mentale Zustände funktionale Zustände sind, die auf verschiedene Arten realisiert werden können. Beim Menschen sind solche funktionalen Zustände faktisch durch Gehirnzustände realisiert. Funktionale Zustände lassen sich definieren durch kausale Relationen zu einem bestimmten Input, einem bestimmten Output und anderen Zuständen des Systems. Die funktionale Organisation des Systems ist dann identisch mit den so definierten Zuständen.

Betrachten wir als Beispiel für eine funktionale Analyse Mausefallen. Mausefallen können physikalisch verschieden realisiert werden, der Begriff "Mausefalle" aber sagt, was alle gemeinsam haben (worin sie funktional äquivalent sind): jede Falle nimmt als Input eine lebendige Maus und liefert als Output eine tote. Genauer: Ist die Mausefalle im Zustand "Bereit", so führt ein Input "lebendige Maus" zu einem Output "tote Maus" und zum Mausefallenzustand "Nicht bereit". Dies funktionale Sichtweise gilt für jede Mausefalle, egal nach welchem Mechanismus sie arbeitet. Analog läßt sich das Programm eines Computers funktional beschreiben. Ist der Computer im Zustand Z, dann führt ein Input I zu einem Output O und zu einem weiteren Zustand Z'. Diese Beschreibung eines abstrakten Automaten kann in der Welt durch verschiedene physikalische Systeme realisiert werden. Das gleiche Programm kann auf physikalisch völlig unterschiedlichen Computern realisiert werden - mechanisch, elektrisch, digital oder analog.

Putnam erläutert am Beispiel der TURING-Maschine den Unterschied zwischen funktionaler und physikalischer Betrachtungsweise, mit der die grundlegende Idee des Funktionalismus vielleicht deutlicher wird (vgl. hierzu Putnam 1985, Beckermann 1985b).

Die TURING-Maschine geht auf den englischen Mathematiker A. Turing zurück, der uns weiter unten noch beschäftigen wird. Eine seiner wichtigsten Leistungen war die erste mathematisch ausgereifte Computertheorie, in der ein spezieller Automat - eben jene TURING-Maschine - erfunden wurde. Diese ist jedoch lediglich ein theoretisches Modell und wurde nie praktisch gebaut.
Eine TURING-Maschine ist ein äußerst einfacher Computer, der aus 2 Teilen besteht: einem eindimensionalen Band und einem Kopf. Das Band ist nur ein passives Speichermedium, das in Felder unterteilt ist, von denen jedes genau 1 Zeichen eines Alphabets enthält. Der Kopf ist der aktive Teil der Maschine. Er kann das Band jeweils um ein Feld nach links oder rechts bewegen und jedes Feld lesen, beschreiben, löschen oder das Band auch stoppen. Außerdem ist der Kopf bei jedem Schritt in einem bestimmten Zustand, der sich in der Regel von Schritt zu Schritt verändert (Abb.1).

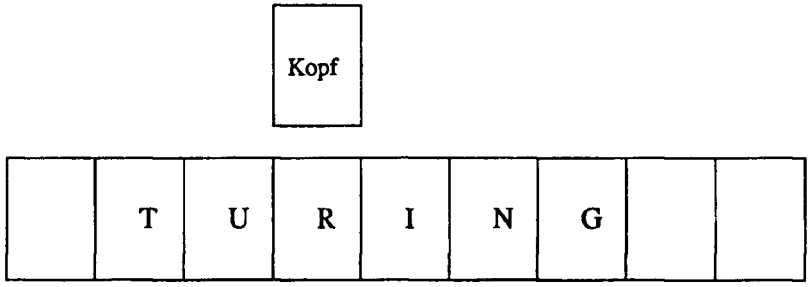

Abb. 1: Schema einer TURING-Maschine.

Was die Maschine tut, hängt vom Zustand des Kopfes ab und dem vorgefundenen Zeichen eines Feldes. Die Arbeitsweise der TURING-Maschine läßt sich damit vollständig durch eine Tabelle angeben.
Betrachten wir zum besseren Verständnis folgendes Beispiel einer TURING-Maschine, die wir mit M bezeichnen (Abb.2).

		Zustände	
		1	2
	-	_ R 2	STOP
	A	A L 1	B R 2
Zeichen	B	B L 1	C R 2
	C	C L 1	A R 2

Abb. 2: Beispiel einer einfachen TURING-Maschine M.

Die *Zustände* der TURING-Maschine M in Abb.2 sind in der Kopfzeile aufgeführt, das *Alphabet* in der linken Spalte. Jedem Zustand und jedem Zeichen ist eindeutig ein Eintrag in der Tabelle zugeordnet. Dieser Eintrag besteht entweder aus drei Zeichen oder dem Wort "STOP". Das erste Zeichen des Eintrags ist das Zeichen, welches geschrieben wird, das zweite gibt an, ob sich der Kopf nach links (L) oder rechts (R) bewegen soll und das dritte gibt den nächsten Zustand (1 oder 2) an.
In unserem Beispiel enthält das Alphabet 4 Zeichen: das Leerzeichen (symbolisiert durch '_') und die Buchstaben A, B, C. Diese Tabelle kann man auch als Verhaltensgesetze auffassen, die beschreiben, was die Maschine tut. Jeder Zelle entspräche ein Gesetz, die Tabelle ließe sich also auch durch 9 Gesetze charakterisieren. Der Zeile 2, Spalte 1 entspräche das Gesetz: Wenn die Maschine im Zustand 1 ist und A eingelesen wird, dann schreibe A, bleibe im Zustand 1 und gehe 1 Feld nach links.

Nehmen wir nun an, das Band ist wie in Abb. 3 beschrieben:

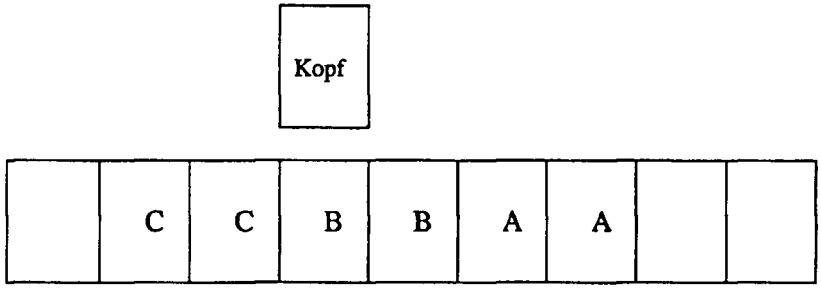

Abb. 3: TURING-Maschine bei Beginn des Programms.

Beginnt die Maschine im Zustand 1 und steht der Kopf über B, so wird Zeile 3, Spalte 1 ausgelöst. Der Kopf schreibt B (also den gleichen Buchstaben nochmal), wandert um eine Stelle nach links und bleibt im Zustand 1. Im nächsten Schritt wird C gelesen, C geschrieben, um eine Stelle nach links gegangen und die Maschine befindet sich weiter im Zustand 1. Dies wird - laut Tabelle -solange fortgesetzt, bis ein Leerzeichen gelesen wird. In diesem Fall (Zeile 1, Spalte 1), wird nichts geschrieben, aber der Kopf wandert nach rechts und geht in Zustand 2 über. Nun wird wieder C gelesen, da aber nun die Maschine im Zustand 2 ist, wird A ausgegeben. Aus der Tabelle ist ersichtlich, daß nun alle C in A, alle B in C und alle A in B umgewandelt werden. Die Maschine stoppt, sobald das erste Leerzeichen gelesen wird. Nachdem dieses - zugegeben recht einfache - Programm abgearbeitet ist, sieht das Band also wie in Abb. 4 aus.

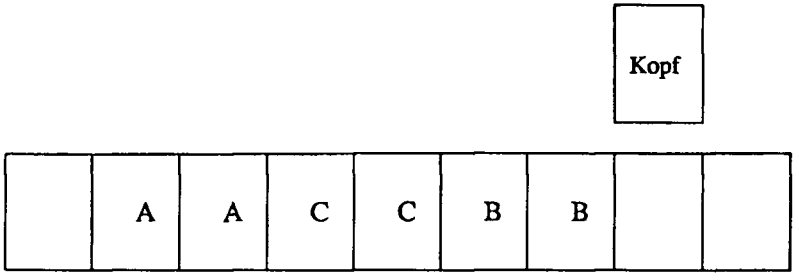

Abb. 4: TURING-Maschine nach Abarbeitung des Programms.

Die zwei Zustände 1 und 2, welche Putnam logische Zustände nennt, werden allein durch ihre funktionalen Eigenschaften charakterisiert, d.h. durch das, was die Maschine M bei gegebenem Input (Bandzeichen) tut, wenn sie sich in einem der Zustände befindet. Die Verhaltensgesetze lassen sich deshalb als implizite funktionale Definitionen dieser

Zustände auffassen. Nun kann aber *jedes* physikalische System P, das die für eine TURING-Maschine erforderlichen Teile besitzt, als eine solche M-Maschine aufgefaßt werden. Vorausgesetzt muß werden, daß sich an P zwei physikalische Zustände unterscheiden lassen, welche die für die logischen Zustände 1 und 2 charakteristischen funktionalen Eigenschaften haben. Tut sie dies, dann arbeitet sie genauso wie M.
Damit kann jede konkrete Maschine M aber auf 2 verschiedene Weisen beschrieben werden. Zum einen ist sie als eine *konkrete* Maschine ein physikalisches System, dessen Zustände mit physikalischen Begriffen beschrieben werden können. Zum andern ist sie eine *abstrakte* Maschine, deren Zustände bestimmte funktionale Eigenschaften haben. Jeder konkreten Maschine sind somit physikalische als auch funktionale Zustände zuzusprechen, die einander entsprechen. Ist P eine M-Maschine, so bedeutet also "P befindet sich im funktionalen Zustand 2" auf physikalischer Ebene "P befindet sich im physikalischen Zustand P2" (vgl. Beckermann 1985b).

Dem Funktionalismus zufolge ist der Begriff der funktionalen Eigenschaft oder Organisation auch auf den menschlichen Geist übertragbar. Danach sind geistige Zustände funktionale Zustände des Körpers, die sich zu den physiologischen Zuständen genauso verhalten wie die logischen Zustände einer Maschine zu ihren physikalischen Zuständen:

Geistiger Zustand		Logischer Zustand
----------	wie	----------
Physiologischer Zustand		Physikalischer Zustand
Mensch		Maschine

Die psychologische Verfassung eines Menschen kann als Teil seiner funktionalen Organisation aufgefasst werden. Hat jemand irgendeine Überzeugung, so ist dies gleichwertig mit der Aussage, daß man etwas über die funktionale Organisation dieses Menschen weiß. Jeder mentale Zustand ist durch die kausale Rolle bestimmt, die er als funktionaler Zustand in der funktionalen Organisation eines Organismus spielt. Im Unterschied zum Behaviourismus erkennt der Funktionalismus somit die kausale Rolle, welche mentale Zustände im Geistesleben spielen, ausdrücklich an (Heyer 1988a).

Dem Funktionalismus unterliegt also weder eine Trennung von Mentalem und Körperlichem wie im Descartes'schen Dualismus, noch eine Gleichsetzung von mentalen Zuständen mit physiologischen oder eine Verleugnung derselben wie im Behaviourismus. Der psychische Zustand eines Systems hängt nicht davon ab, ob das System ein Mensch oder eine Maschine ist. Es kommt lediglich darauf an, auf welche Weise diese Systeme zusammengesetzt und organisiert sind. Es könnte Systeme mit genau derselben psychologischen Verfassung wie Menschen geben, die aber eine ganz andere physikalische oder physiologische Konstitution hätten. Zwar sind beim Menschen mentale Zustände faktisch an Gehirnzustände gebunden, die kausal wirksam sind. Der Funktionalismus

schließt aber die Möglichkeit nicht aus, daß Maschinen denken und fühlen können, auch wenn diese Möglichkeit absurd erscheinen mag (Foder 1981). Jedes beliebige System mit dem richtigen Programm, den richtigen Inputs und Outputs hätte dann im gleichen Sinn einen Geist, in dem Menschen einen Geist haben.

Ein mit dem Funktionalismus eng verknüpfter Aspekt ist die hierarchische Organisation von Systemen, die es erlaubt, komplexere Aufgaben auf immer einfachere Teilaufgaben zu reduzieren. Dies führt dazu, daß die Systeme auf verschiedenen Ebenen beschreibbar sind. Auf der obersten Ebene spricht man von Schachcomputern etwa wie von Systemen, die über bestimmte Informationen verfügen, bestimmte Absichten verfolgen und darauf basierend bestimmte Handlungen vornehmen. (Übrigens haben Benutzer eine Tendenz, dies nicht nur mit intelligenten Schachprogrammen zu machen, sondern bereits mit einfachen Textverarbeitungssystemen). Dennett nennt dies die Ebene der intentionalen Einstellung. Diese intentionale Ebene kann auf die Programmebene reduziert werden: das, was der Schachcomputer tut (intentionale Ebene) korrespondiert bestimmten Anweisungen im Programm. Kennt man das Programm genau, so kann man jeden Zug des Computers vorhersagen. Schließlich läßt sich nochmal eine Reduktion auf die physikalische Ebene durchführen. Hier fließen Ströme und Voraussagen stützen sich auf physikalische Zustände des Systems unter Anwendung des Wissens über Naturgesetze.

In KI-Systemen werden nun intentionale Zustände durch physikalische Zustände realisiert. Intentionale Einstellungen lassen sich vollständig auf die physikalische Ebene reduzieren. Übertragen auf den menschlichen Geist bedeutet dies nach funktionalistischer Auffassung, daß dieser ein Konstrukt ist, der aus weniger intelligenten Sub-Systemen besteht. Jedes dieser Sub-Systeme nimmt bestimmte Funktionen wahr und delegiert an noch weniger intelligente Sub-Sub-Systeme usw.(Heyer 1988a).

Fassen wir die wichtigsten Konsequenzen des Funktionalismus für die KI zusammen, die als die Arbeitshypothesen der heutigen, kognitiv ausgerichteten KI betrachtet werden können und von vielen - wenn auch nicht allen - KI-Forschern akzeptiert werden.

- Für die Beschreibung der Organisation eines Systems muß die tatsächliche physikalische Realisierung nicht bekannt sein. Details, ob das System elektronisch ist oder physiologisch oder hydraulisch oder sonstwie sind völlig belanglos. Mensch und Computer gehören zu einer gemeinsamen Gattung informationsverarbeitender Systeme mit dem - unwichtigen - Unterschied, daß die physikalische Trägersubstanz beim Computer Hardware, beim Menschen Gehirn-"Wetware" ist. Am Geist ist nichts wesentlich biologisches.
- Die funktionale Architektur des menschlichen Geistes und die des Computers sind vergleichbar. Auf beiden Trägersubstanzen können Organisation und Zusammenspiel der einzelnen funktionalen Zustände einander entsprechen.
  Menschen sind entweder im Prinzip Maschinen oder Maschinen sind im Prinzip zu den gleichen kognitiven Leistungen fähig wie Menschen. Bei geeigneter Programmierung können Computer somit menschliche kognitive Fähigkeiten nachvollziehen.

- Wenn sich in Menschen und Computern dieselben theoretischen Strukturen realisieren, die fähig sind, Symbolverarbeitung durchzuführen, so lassen sich Korrespondenzen herstellen zwischen abstrakten Software-Strukturen in Maschinen und postulierten abstrakten Mentalstrukturen in Menschen. Da wir die Prinzipien von Computerprogrammen verstehen, nicht jedoch die Prinzipien menschlichen Verhaltens und Denkens, können mit Hilfe der Computerprogramme Rückschlüsse von Mechanismen in Computerprogrammen auf Mechanismen in empirischen, menschlichen Systemen vorgenommen werden. Mit anderen Worten: Computerprogramme helfen uns, den menschlichen Geist besser zu verstehen.

4. Die physikalische Symbolsystemhypothese

Wenn Mensch und Computer Instanzen einer abstrakten Gattung informationsverarbeitender Systeme sind, so müssen die elementaren Prozesse der Informationsverarbeitung (Empfang, Analyse, Rekonstruktion, Speicherung, Ausgabe) einander entsprechen. Wie lassen sich nun solche informationsverarbeitenden Systeme allgemein beschreiben und zu welchen Leistungen sind sie fähig?

Newell und Simon (1976) definieren ein solches System - sie nennen es physikalisches Symbolsystem - als eine Menge von Symbolen, das sind physikalische Muster, Zeichenketten wie "ABC" oder "HANS". Symbole setzen sich zu Symbolstrukturen oder Ausdrücken zusammen, wie etwa "HANS GEHT SCHWIMMEN". Zwischen den Symbolen bestehen bestimmte Beziehungen, z.B. liegen "GEHT" und "SCHWIMMEN" nebeneinander. Neben solchen Symbolstrukturen enthält das System Prozesse, die auf diesen Symbolstrukturen operieren um weitere Strukturen zu erzeugen. Die entsprechenden Prozesse können etwa sein Erzeugungs-, Abänderungs-, Reproduzierungs- und Zerstörungsprozesse. Ein Zerstörungsprozeß könnte aus der obigen Symbolstruktur etwa "HANS GEHT" machen, ein Abänderungsprozeß "WILLY GEHT SCHWIMMEN".

Ein physikalisches Symbolsystem ist dann eine Maschine, die eine sich entwickelnde Sammlung von Symbolstrukturen erzeugt.

Die physikalische Symbolsystemhypothese macht nun eine Aussage über die Fähigkeiten solcher Systeme; sie lautet:

Ein physikalisches Symbolsystem besitzt hinreichende und notwendige Mittel für eine allgemeine intelligente Tätigkeit (Newell/Simon 1976).

Diese Aussage kann nicht analytisch-formal bewiesen oder widerlegt werden. Vielmehr handelt es sich um eine empirische Behauptung, die durch empirische Forschung verifiziert oder falsifiziert werden kann.

Die Hypothese macht Aussagen in 2 Richtungen:

Erstens sind die Mittel eines physikalischen Symbolsystems hinreichend. Dies bedeutet, daß die Mittel des Systems ausreichen, um intelligente Handlungen erzeugen zu können. Hierauf deuten bereits existierende Programme und Erkenntnisse der KI hin.

Zweitens sind die Mittel notwendig, d.h. intelligentes Handeln erfordert ein physikalisches Symbolsystem. Evidenz liefert der informationsverarbeitende Ansatz in der kognitiven Psychologie.

Die Hypothese wird nicht zuletzt durch negative Evidenz gestützt. Es gibt keinen konkurrierenden Ansatz, der Aufschluß darüber geben würde, wie intelligentes Handeln - beim Menschen oder bei der Maschine - erreicht werden könnte.

Damit bildet diese Hypothese den Kern der heutigen KI:

"Die physikalische Symbolsystemhypothese ist in zweifacher Hinsicht wichtig. Sie ist eine signifikante Theorie der Natur menschlicher Intelligenz und ist daher von großem Interesse für Psychologen. Außerdem bildet sie die Grundlage für den Glauben, daß es möglich ist, Programme zu erstellen, die bis jetzt nur von Menschen durchführbare intelligente Aufgaben durchführen können" (Rich 1988:5).

5. Der TURING-Test

Wenn es das Ziel ist, intelligente Maschinen zu bauen, wie erkennen wir dann, ob wir erfolgreich sind? Schließlich muß es irgendwelche Kriterien geben, an denen die erfolgreiche Bewältigung dieser Aufgabe gemessen werden kann.

Man könnte auf die Idee kommen, psychologische Intelligenztests als Maßstab zu nehmen. Dies ist aber aus zweierlei Gründen nicht möglich: IQ-Tests messen lediglich den Grad der Intelligenz, setzen aber bereits eine bestimmte Intelligenz voraus. Bei Computern ist aber gerade das Problem, ob ihnen überhaupt Intelligenz zugesprochen werden kann. Zum andern sind IQ-Tests auf Menschen zugeschnitten. Was gebraucht wird, ist eine allgemeiner Test, ob ein Ding überhaupt intelligent ist (Haugeland 1987: 6).

Der Mathematiker Alan Turing schlug 1950 ein Imitationsspiel vor, das ein solches - allgemeines - Kriterium liefert. In seinem Aufsatz "Computing Machinery and Intelligence" (deutsche Übersetzung "Maschinelle Rechner und Intelligenz" in Hofstatter/Dennett 1981) geht er von der Frage aus, ob Computer jemals werden denken können. Turing, angewidert von fruchtlosen Diskussionen um diese emotional befrachtete Frage, ersetzt diese durch eine andere, unzweideutige Frage, welche einen operationalen Test dieses Problems darstellt. Dieser Test, der in die Literatur als TURING-Test eingegangen ist, stellt die neue Fassung des Problems "Können Maschinen denken" dar. Nach der Originalversion von Turing läßt sich der Test als folgendes Spiel beschreiben (vgl. Turing 1981):

An dem Spiel nehmen drei Personen teil: ein Mann (A), eine Frau (B) und ein Fragesteller (C), dessen Geschlecht keine Rolle spielt. Der Fragesteller befindet sich in einem eigenen Raum, getrennt von den beiden anderen Personen. Der Fragesteller hat die Aufgabe herauszufinden, wer von den beiden anderen Personen der Mann und die Frau ist. Der Fragesteller kennt sie lediglich unter der Signatur X und Y. Ist das Spiel zu Ende, dann sagt er entweder "X ist A und Y ist B" oder umgekehrt "X ist B und Y ist A". C darf Fragen beliebiger Art stellen, die Aufgabe des Mannes A ist es, den Fragesteller möglichst dahin zu bringen, eine falsche Identifizierung vorzunehmen, während die Frau B das Ziel

verfolgt, dem Fragesteller zu helfen. Damit der Fragesteller die beiden anderen nicht an der Stimme erkennt, erfolgt der Austausch der Antworten über beschriebene Zettel bzw. getippt (Turing spricht von Fernschreiberverbindungen, heute würde man wohl Bildschirme benutzen).

Wie werden sich A und B wohl verhalten? B wird wahrscheinlich am besten ehrlich antworten, etwa mit Bemerkungen wie "Ich bin eine Frau, hören Sie nicht auf ihn!" aber das hilft nichts, da der Mann ebenfalls Äußerungen dieser Art machen wird.

Wird es dem Fragesteller gelingen, eine richtige Identifizierung vorzunehmen? Hinreichendes Intelligenzniveau von A und B vorausgesetzt, wird sich der Fragesteller sicherlich schwer tun. Bei einem statistischen Test mit 100 Versuchspersonen und unter der Prämisse, daß Indifferenz nicht zugelassen ist, würde man erwarten, daß ca. 50 Personen X mit A und Y mit B identifizieren und die anderen 50 die umgekehrte Identifikation vornehmen.

Was hat dies alles mit der Frage zu tun, ob Computer denken können?

Ganz einfach. Turing ersetzt den Mann durch eine Maschine, die die gleiche Rolle spielen muß. Die entscheidende Frage ist nun: kommt der Fragesteller jetzt auch genauso oft zu einem falschen Ergebnis wie im Originalspiel? Genau diese Frage tritt an die Stelle der ursprünglichen Frage "Können Maschinen denken?" Wird sie positiv beantwortet, so ist unbestreitbar der Maschine Intelligenz zu attestieren.

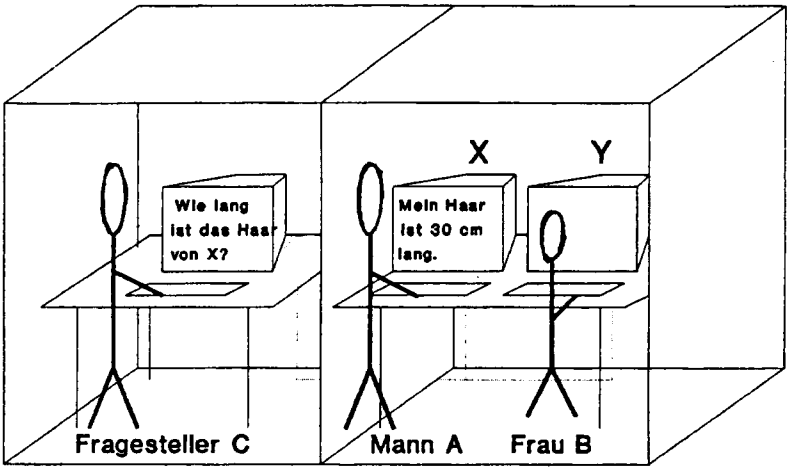

Abb. 5: Bildliche Veranschaulichung des TURING-Tests I.

Der Fragesteller C interviewt einen Mann A und eine Frau B, die er lediglich unter der Signatur X und Y kennt. Der Fragesteller muß mit beliebigen Fragen herausfinden, wer der Mann und wer die Frau ist.
C könnte also eintippen:
> Kann X mir sagen, welche Kleidung er oder sie heute trägt?
Nehmen wir an, X ist der Mann A. Da dieser versuchen wird, C dahin zu bringen, eine falsche Identifizierung vorzunehmen, könnte er etwa antworten:
> Ich trage ein geblümte Bluse, eine Perlenhalskette, einen halblangen schwarzen Rock und schwarze Schuhe.
Diese Antwort erscheint auf dem Bildschirm im Zimmer von C und C kann die nächste Frage eingeben.

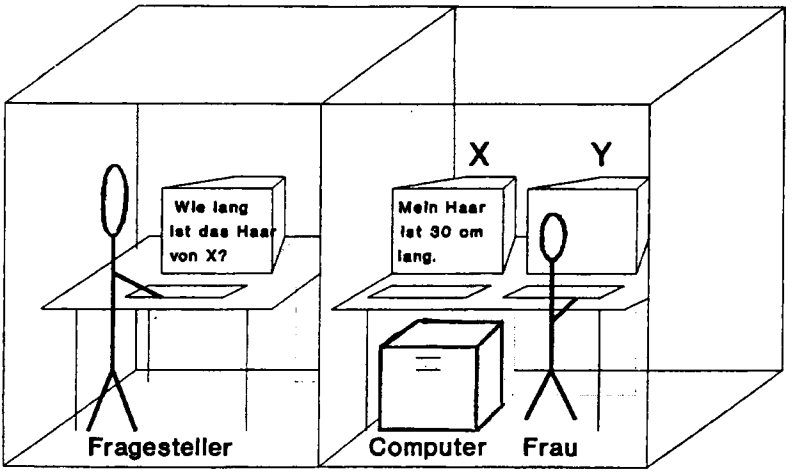

Abb. 6: Bildliche Veranschaulichung des TURING-Tests II.

Beim TURING-Test wird der Mann durch eine Maschine ersetzt. Kommt der Fragesteller genauso oft zu einem falschen Ergebnis wie beim Originalspiel, so hat die Maschine den Intelligenz-Test bestanden.

Die Essenz dieses Tests ist vielleicht für jemanden, der erstmals damit konfrontiert ist, nicht ohne weiteres einsehbar. Wesentlich ist nicht der Bildschirm oder das Täuschungsmanöver, was nur dazu dient, dem ganzen einen experimentellen Anstrich zu geben. Worauf es in diesem Test wirklich ankommt ist: kann die Maschine einen Menschen so überzeugend nachahmen, daß diese von einem Menschen nicht mehr zu

unterscheiden ist? Ausschlaggebend ist hierbei das Reden und das Wissen: Redet die Maschine (über das Terminal) wie ein Mensch und hat die Maschine auch das Wissen eines Durchschnittsmenschen? Die Maschine muß nämlich ungefähr das sagen, was durchschnittliche Menschen in solchen Situationen sagen würden. Genauso muß die Maschine über Themen Bescheid wissen, worüber Menschen normalerweise sprechen: Politik, Sonette, den Wetterbericht, Liebesfilme, Gott, den Dollarkurs. Könnte eine Maschine antworten wie Menschen das üblicherweise tun, so würde sie nicht nur die deutsche Sprache exzellent beherrschen, sondern es müßte ihr auch ein großartiges Verständnis für Gedichte, die Jahreszeiten, menschliche Gefühle etc. bescheinigt werden. Das Geniale an dem Test ist, daß Turing's Frage-Antwort-Spiel jeden menschlichen Bereich einbringen kann, den der Fragesteller einbringen will.

Natürlich müßte die Maschine den Fragesteller auch in Aufgaben täuschen, die sie gut beherrscht, sie aber als Maschine verraten würde. Wenn C frägt "Wieviel ist 34894823 * 6546156126" so könnte sie in Sekundenbruchteilen die richtige Antwort geben; sie wird sich aber so verhalten, wie Menschen sich verhalten und einige Zeit brauchen, bis sie das Ergebnis - das vielleicht auch falsch sein kann - ausgibt.

Turing's Test zeigt auch sehr schön, wie Intelligenz ohne Bezugnahme auf eine physikalische Trägersubstanz geprüft werden kann. Er abstrahiert völlig von der physikalischen Realisation der Intelligenz. Intelligenz ist nicht an die biologische Trägermasse Gehirn gebunden und es würde nichts bringen, eine Denkmaschine dadurch menschlicher zu machen, daß sie in künstliches Fleisch eingebettet wird. Unwichtige physische Eigenschaften - Aussehen, Material, Stimme - werden durch die Versuchsanordnung ausgeschaltet, erfaßt wird das "reine" Denken. Damit können KI-Forscher Nebensächlichkeiten völlig ausschalten und sich den grundlegenden theoretischen Fragen zuwenden.

Turing diskutiert in seinem Aufsatz eine Reihe von Einwänden gegen seinen Test als Intelligenzkriterium für Maschinen, die er alle entkräftigt. Einer der interessantesten Einwände stammt von Lady Ada Lovelace, jener Dame, die Babbage's Analytische Maschine mitentwickelte und das erste Computerprogramm schrieb.

Charles Babbage (1792-1871) war ein etwas exzentrischer Mathematiker, der im Laufe seines Lebens einen wahnhaften Haß auf Werkzeugmacher und Straßenmusikanten entwickelte. 1822 entwirft der kauzige Babbage die Analytische Maschine, einen Rechenautomaten, der alle wesentlichen Funktionen heutiger Computer beinhaltete. Die Analytische Maschine wurde von Babbage zwar nie erbaut, sie gilt aber als Urbild heutiger Computerarchitekturen und wäre uns unbekannt, wenn es nicht Gräfin Ada Lovelace gegeben hätte. Lady Lovelace war mit Babbage befreundet und von seiner Analytischen Maschine begeistert. Im Rahmen einer Arbeit über Babbage's Maschine entwickelte sie die ersten Computerprogramme. Lady Ada wurde darüberhinaus in zweierlei Hinsicht berühmt: erstens wurde eine Programmiersprache nach ihr benannt und zweitens stammt von ihr der vielzitierte Ausspruch, wonach Computer nur das tun können, wozu sie programmiert sind.

Der Einwand lautet sinngemäß: Computer können nicht schöpferisch (und damit intelligent) sein, weil sie nur das tun können, wozu sie programmiert sind.

Die folgenden Gedankengänge zur Entkräftigung des Lady Lovelace Einwands beruhen auf Turing und wurden von Haugeland (1987: 8-9) detaillierter ausgearbeitet.

Haugeland bemerkt, daß es in einem technischen Sinn sicherlich richtig ist, daß Computer lediglich Programme ausführen. Dies beweist aber gar nichts, denn einen ähnlichen Einwand könnte man ebenfalls in Bezug auf Menschen vorbringen:

- Angenommen, wir wissen im Detail um die Prozesse im Gehirn und kennen alle neurophysiologischen Gesetze. Dann agieren unsere Gehirne gemäß diesen Gesetzen und wir wären in der Lage, Vorhersagen darüber zu machen, was wir tun oder denken werden. Obwohl unser Denken "programmiert" ist, sich gemäß diesen Gesetzen zu verhalten, wären wir aber weiterhin kreativ.
- Zur Kreativität programmiert sein, ist kein Widerspruch in sich, denn auch Menschen basieren auf einem Entwurf, nämlich auf dem Ergebnis der Evolution. Normalerweise tun wir das, wozu wir entworfen sind und die Kreativität ist ein Teil unseres evolutionären Entwurfs.
- Allerdings läßt sich einwenden, daß dies nur eine Metapher ist, da die Evolution kein realer Konstrukteur ist, sondern ein unbeseelter Naturvorgang. Computer werden dagegen im wörtlichen Sinn programmiert. Wenn ein Computerprogramm kreativ ist, handelt es sich um die Kunstfertigkeit des Programmierers und nicht der Maschine.

Aber warum sollte Kreativität durch Abstammung determiniert sein? Ob ein kreatives Programm von einem Programmierer erstellt wird oder aus einer Panne im Labor hervorgeht, ist für die Leistungsfähigkeit des Programms völlig unwichtig. "Was direkt programmiert wird, ist nur ein Bündel allgemeiner Informationen und Prinzipien, nicht viel anders als das, was Lehrer ihren Schülern eintrichtern. Was danach geschieht, was also das System mit diesem ganzen Input macht, kann der Konstrukteur (oder der Lehrer oder sonstwer) nicht vorhersagen. Die eindrucksvollsten Beispiele sind Schachcomputer, die besser als ihre Programmierer spielen und mit brillanten Zügen aufwarten, auf die letztere nie gekommen wären" (Haugeland 1987: 9).

Ähnlich wie Haugeland argumentiertenSchank/Childers:

"Die Tatsache, daß alles, wozu man einen Computer bringen kann, immer "nur ein Programm" ist, setzt ihn nicht herab. Auch ein Mensch ist in gewisser Weise nur ein Programm. In dem Maße, wie das Verhalten eines Menschen jederzeit eine Funktion seiner Erfahrung darstellt, ist er indirekt von seinen Erfahrungen *programmiert* (oder direkt von den Eltern oder der Schule). Was wir im Moment tun, ist im wesentlichen von dem beeinflußt, was wir früher erlebt haben. Das bedeutet nicht, daß jeder, wenn er denselben Erfahrungen ausgesetzt gewesen wäre, sich genauso verhalten würde oder sollte. Menschen sind verschieden - sowohl nach ihrer angeborenen Intelligenz wie auch ihrer Erfahrung. Eben *weil* wir in gewissem Sinne von unserer Erfahrung *programmiert* sind, verhalten wir uns alle unterschiedlich und mit so unglaublicher Vielfalt" (Schank/Childers 1986: 69, Hervorhebungen von Schank/Childers).

Der TURING-Test wird von den meisten kognitiv orientierten KI-Forschern als Intelligenzkriterium für Maschinen akzeptiert. Turing selbst war überzeugt, daß eine geeignet programmierte Maschine bis zum Ende des Jahrhunderts seinen Test bestehen würde. Unabhängig davon war er überzeugt, daß sich bis dahin der Wortgebrauch so verändert hat, daß man von Maschinen als denkend wird sprechen können.

Es bedarf keiner besonderen Betonung, daß heutige Rechner beim TURING-Test massiv versagen. Selbst die intelligentesten natürlichsprachlichen Programme haben nicht die geringste Chance, diesen Test zu bestehen. Von dem oberflächlich eleganten und sehr intelligent erscheinenden ELIZA (siehe Kasten) würde man vielleicht erwarten, daß es bei der Eingabe von

> Ich fühle mich heute Blablabla

vielleicht anwortet mit

> Was meinen sie mit Blablabla

Stattdessen antwortet es ungefähr mit

> Es tut mir leid von Ihnen zu hören, dass Sie sich BlaBlaBla fühlen.

ELIZA ist ein frühes Computerprogramm von Joseph Weizenbaum, das einen nicht-direktiven Gesprächstherapeuten nach der Methode von Rogers simuliert. Danach spiegelt der Therapeut lediglich die Antworten seines Klienten mit eigenen Worten wider, ohne aber selbst die Intitiative zu übernehmen oder eigene Gedanken bzw. Vorschläge einzubringen. Der Patient wird dadurch zum Sprechen gebracht, daß ihm seine eigenen Äußerungen wie bei einem Echo zurückgegeben werden. Das wiedergegebene Gesprächsprotokoll ist ein Ausschnitt aus einem Interview, das eine junge Frau mit ELIZA führte (Weizenbaum 1978: 15-16).

Mensch	>	Die Männer sind alle gleich.
Maschine	>	In welcher Weise
Mensch	>	Sie machen uns immer wegen irgendwas verrückt
Maschine	>	Können Sie ein Beispiel nennen
Mensch	>	Naja, die Idee zu diesem Gespräch stammt von meinem Freund.
Maschine	>	Die Idee zu diesem Gespräch stammt von Ihrem Freund
Mensch	>	Er sagt, dass ich oft deprimiert bin.
Maschine	>	Es tut mir leid von Ihnen zu hören, dass Sie deprimiert sind
Mensch	>	Das stimmt, ich bin unglücklich.
Maschine	>	Glauben Sie, dass dieses Gespräch Ihnen hilft, nicht mehr unglücklich zu sein

Mensch	>	Ich brauche jedenfalls Hilfe, das steht fest.
Maschine	>	Was würde es für Sie bedeuten, Hilfe zu bekommen
Mensch	>	Vielleicht könnte ich lernen, mit meiner Mutter besser auszukommen.
Maschine	>	Erzählen Sie mir mehr über Ihre Familie
Mensch	>	Meine Mutter erledigt alles für mich.
Maschine	>	Wer in Ihrer Familie erledigt noch alles für Sie
Mensch	>	Mein Vater.
Maschine	>	Ihr Vater
Mensch	>	Sie sind ein wenig wie mein Vater.
Maschine	>	Welche Ähnlichkeit sehen Sie
Mensch	>	Sie sind nicht besonders aggressiv, aber ich glaube, Sie möchten nicht, dass ich das beachte.
Maschine	>	Was lässt Sie vermuten, dass ich nicht sehr aggressiv bin
Mensch	>	Sie streiten nicht mit mir.
Maschine	>	Wieso glauben Sie, dass ich nicht mit Ihnen streite
Mensch	>	Sie haben Angst vor mir.
Maschine	>	Macht Ihnen der Gedanke Spass, dass ich vor Ihnen Angst habe
Mensch	>	Mein Vater hat vor jedem Angst.
Maschine	>	Was fällt Ihnen noch ein, wenn Sie an Ihren Vater denken
Mensch	>	Angeber.
Maschine	>	Hat das etwas damit zu tun, dass die Idee zu diesem Gespräch von Ihrem Freund stammt

Mancher Leser, dem dieses Beispiel noch nicht bekannt ist, mag vielleicht überrascht sein, daß Maschinen Dialoge dieser Art führen können und zu welch "einfühlendem" Verständnis Computer heutzutage fähig sind. Leider versteht ELIZA in Wirklichkeit überhaupt nichts. Es täuscht Verstehen lediglich vor und weiß weder über Väter oder Mütter, noch über depressive Gefühlszustände noch irgendeinen anderen menschlichen Gegenstandsbereich Bescheid.

Das Programm enthält eigentlich nur festgespeicherte Satzbausteine (Schablonen) und einige Regeln. Eingabesätze werden an diese Schablonen angepaßt und jeder dieser Eingabeschablonen sind eine oder mehrere Ausgabeschablonen zugeordnet, die zur Antwortgenerierung benutzt werden. ELIZA ist ein klassisches Beispiel dafür, wie mit einfachen Methoden des sog. Pattern Matching Verstehen vorgetäuscht werden kann.

Der oben wiedergegebene Dialog ist die deutsche Übersetzung des englischen Originals. Aufgrund der Besonderheiten der deutschen Sprache ist es schwierig, das Programm ins Deutsche zu übertragen. Dem Verfasser ist kein dem englischen Original vergleichbare deutsche Fassung des Programms bekannt.
Noch eine kurze Bermerkung dazu, was ELIZA ausgelöst hat.
Weizenbaum war über die Wirkungen seines Programms entsetzt.
Praktizierende Psychiater glaubten, daß das Programm zu einer automatischen Form von Psychotherapie ausgebaut werden könnte. Kenneth Colby, ein Psychiater aus Stanford, von dem später noch die Rede sein wird, meinte, das Programm könne als therapeutisches Werkzeug dienen. Weizenbaum wehrte sich vehement gegen diesen Mißbrauch seines Programms, für das es nicht gedacht war.
Aber auch Laien waren fasziniert von ELIZA. Personen, die sich mit dem Programm unterhielten, bauten eine emotionale Beziehung zu der Maschine auf und schrieben ihr menschliche Eigenschaften zu. Es kam oft vor, daß man Weizenbaum um Erlaubnis bat, sich mit dem Programm unbeobachtet unterhalten zu dürfen und trotz seiner Erklärungen bestanden die Personen nach der Unterhaltung darauf, das System hätte sie wirklich verstanden.
Dies verdeutlicht die folgende Anekdote:
"Einmal führte meine Sekretärin eine Unterhaltung mit ihm (ELIZA, Anm.d.Verf.); sie hatte seit Monaten meine Arbeit verfolgt und mußte von daher wissen, daß es sich um ein bloßes Computerprogramm handelte. Bereits nach wenigen Dialogsätzen bat sie mich, den Raum zu verlassen. Ein andermal äußerte ich die Absicht, das System so zu schalten, daß man alle Unterhaltungen abrufen konnte, die z.B. in einer Nacht mit ihm geführt worden waren. Sofort wurde ich mit Vorwürfen überschüttet, mein Vorschlag laufe darauf hinaus, die intimsten Gedanken anderer auszuspionieren; ein deutliches Anzeichen dafür, daß sich die einzelnen mit dem Computer unterhalten hatten, als sei er eine Person, der man sich in geeigneter und sinnvoller Weise über Privatangelegenheiten mitteilen konnte" (Weizenbaum 1978:19).
Weizenbaum selbst wurde aufgrund der Erfahrungen mit seinem Programm zu einem KI-Skeptiker und -kritiker, arbeitete aber weiter auf diesem Gebiet. Sein populäres Buch "Die Macht der Computer und die Ohnmacht der Vernunft" ist eine sehr lesenswerte, locker geschriebene und kritische KI-Einführung, ohne in technische Details zu gehen.

Und mit dieser Antwort hat die Maschine den Turing-Test bereits nicht bestanden.

Trotzdem ist die grundlegende Idee des Turing-Tests derzeit nicht so unnütz, wie sie vielleicht erscheinen mag. Wir können uns nämlich mit weniger zufrieden geben als der vollständigen Imitation einer Person, indem wir nur bestimmte beschränkte Gebiete beurteilen lassen. Es muß dann geprüft werden, ob das Programmverhalten dem menschlichen Verhalten auf diesem relevanten Gebiet entspricht.

Ganz allgemein kann die Frage, wie wir wissen, ob wir ein intelligentes Programm geschrieben haben, durch die Beurteilung von bereichsspezifischen Experten ersetzt werden. Ein Schachprogramm kann von einem Schachexperten beurteilt werden, der

gegen das Programm spielt. Dies ist ein ziemlich präzises Maß für die Leistung eines Programms. Die Leistungsfähigkeit eines Systems, das medizinische Diagnosen erstellen kann, hängt davon ab, inwieweit es ein Verhalten an den Tag legt, das Experten als Fachurteil eines Kollegen akzeptieren würden - was sicherlich schon weniger präzise ist. Andererseits gibt es Expertensysteme, deren Analysen als Originalforschungsergebnisse publiziert wurden. Solche Systeme arbeiten sicherlich kompetent (Rich 1988:20-21).

Man kann solche Prüfungen auch streng experimentell im Rahmen sogenannter modifizierter TURING-Tests durchführen, bei denen man die ansprechbaren Themen auf bestimmte relevante Bereiche einschränkt.

Modifikationen des TURING-Tests eignen sich insbesondere auch zur Beurteilung von Computermodellen. Eine Variante des Original-Tests besteht z.B. darin, daß man erfahrenen Gutachtern 2 verschiedene Protokolle vorlegt. Das eine Protokoll enthält Interviewaufzeichnungen von einem empirischen System - meist Personen. Das andere Protokoll enthält die Aufzeichnungen des interviewten Computermodells. Wenn die Gutachter - vereinfacht gesagt - nicht in der Lage sind zu identifizieren, welches die Modellausgabe und welches die Ausgabe des empirischen Systems ist, so ist dies ein Hinweis auf die Gültigkeit des Computermodells. Es tut dann nämlich das, was Menschen in ähnlichen Situationen auch tun. Ein berühmtes Programm, das einen solchen Test bestanden hat, ist PARRY, das einen Paranoiker simuliert.

Wir haben oben ELIZA vorgestellt, ein Programm, das einen nicht-direktiven Gesprächspsychotherapeuten simuliert. Der mit dem Computer interagierende Mensch fungierte hier als (neurotischer) Patient, das Programm als Therapeut oder Interviewer. Nun liegt es natürlich auf der Hand, die Rollen zu vertauschen und den Computer den Patienten simulieren zu lassen, während der Mensch diesen interviewen kann.
Genau dies leistet das Programm PARRY des amerikanischen Psychiaters K.M.Colby. PARRY simuliert einen krankhaft paranoiden Patienten, der sich über die Tastatur interviewen läßt. Das Programm ist aber weit mehr als ein bloßes Gegenstück zu ELIZA. Wie oben angedeutet wurde, erzeugt ELIZA seine Ausgaben mit einfachen Mustererkennungsprozeduren, ohne auf die Semantik der Eingabesätze zu achten oder auf der Grundlage irgendeines Modells zu handeln. Im Gegensatz zu ELIZA steckt hinter PARRY aber ein theoretisches Modell der Paranoia, und PARRY verhält sich auf der Grundlage und in Einklang mit diesem Modell. Der Algorithmus von PARRY generiert linguistisches Output-Verhalten, das typisch für Patienten ist, deren Symbolverarbeitung von Paranoia dominiert ist.
Zum Verständnis des Programms ein paar Worte zur Paranoia: Paranoia ist eine psychische Krankheit, die von leichten, kaum merklichen Symptomen bis zu schwerwiegenden Beeinträchtiungen mit Klinikeinweisung reichen kann. Paranoide Patienten haben ein Wahnvorstellungssystem, nach dem sie verfolgt oder bedroht werden. Wahn und Mißtrauen beeinflussen ihre Kommunikation und die Diagnose der Paranoia beruht in erster Linie auf mißtrauischen und feindseligen Antworten auf neutrale Fragen. Die Wahnvor-

stellungen des Patienten können direkt in Form von Bezichtigungen oder Verfolgungsgedanken ausgedrückt werden oder indirekt als Überempfindlichkeit, Sarkasmus, Feindseligkeit, unkooperatives Verhalten etc.
Im Modell ist ein hypothetisches Individuum repräsentiert. Diese Person - nennen wir sie A - ist ein 28-jähriger männlicher protestantischer Börsenangestellter, der allein lebt und seine Eltern selten sieht. A reagiert empfindlich auf Themen, die seine Eltern, seine Religion und sein Geschlechtsleben betreffen. Sein Hobby ist es, bei Pferderennen zu wetten. Einige Monate vorher wurde A in einen Streit mit einem Buchmacher verwickelt und A behauptet, der Buchmacher zahlte eine Wette nicht aus. Nach dem Streit hatte A den Eindruck, daß Buchmacher Schutzgelder an die Unterwelt zahlen und daß dieser Buchmacher dadurch Rache nehmen könnte, daß der Buchmacher auf ihn, A, die Mafia ansetzen oder ihn töten lassen könnte. A ist erpicht darauf, seine Geschichte zu erzählen, um Hilfe zum Schutz vor der Unterwelt zu bekommen. Er beantwortet Fragen über neutrale Themen bereitwillig und gibt Hinweise auf sein Wahnsystem um die Einstellung des Interviewers ihm gegenüber zu ermitteln.
Der Zweck von PARRY war nach Colby, mit einem Informationsverarbeitungsmodell die paranoide Psychose besser zu verstehen. Im Gegensatz zu ELIZA ist PARRY somit ein für die theoretische Psychologie äußerst interessantes Programm.
Darüberhinaus wurde PARRY innerhalb der KI-Welt berühmt, da es einen modifizierten TURING-Test bestand. Es wurde nämlich geprüft, ob Input-Output(I-O)-Paare des Modells zu I-O-Paaren des unterliegenden empirischen Systems ungefähr äquivalent sind. Ist dies der Fall, so ist dies ein Hinweis auf die Gültigkeit des Modells.
Zur Prüfung des Modells führte Colby eine Serie von Tests durch. In der ersten Serie führten Psychiater Interviews mit psychisch Kranken über ein Terminal mit dem Ziel, zu einer Diagnose zu gelangen. Den Ärzten wurde die Befragung per Bildschirm damit erklärt, daß nicht-linguistische Äußerungen wie Stottern oder Zappeln eliminiert werden sollten. In einigen Fällen kommunizierten die Ärzte ohne ihr Wissen nicht mit psychisch Kranken, sondern mit PARRY. Kein einziger Interviewer bemerkte, daß er einen Computer diagnostizierte.
Streng genommen ist diese Variante natürlich kein TURING-Test, wie er oben vorgestellt wurde, da die interviewenden Psychiater nicht beurteilen mußten, welche Antworten von der Maschine und welche von den Patienten kamen. Sie wurden nicht einmal informiert, daß ein Imitationsspiel stattfand. Colby rechtfertigte dies damit, daß die Information, wonach Computer beteiligt sind, die normale Interviewstrategie ändern könnte. Es bestünde dann die Gefahr, daß eher Fragen gestellt werden könnten um herauszufinden, welche Antworten von der Maschine stammen könnten, als relevante Fragen zur Diagnose zu stellen (Boden 1987: 534).
Bei der zweiten Serie von Tests wurde einer anderen Gruppe von Psychiatern die Interviewprotokolle der ersten Serie vorgelegt. Diese Gruppe hatte die Aufgabe, die Protokolle auf Vorliegen oder Nicht-Vorliegen von Paranoia einzustufen; bei Vorliegen, sollte der Grad der Paranoia angegeben werden. Das Ergebnis war, daß Interviews mit der schwachen Version von PARRY als weniger paranoid eingestuft wurden als Interviews mit der starken Version (es gibt eine stark paranoide und eine schwach paranoide Form des Programms).

In der dritten Serie wurden zufällig ausgewählten angesehenen Psychiatern je 2 Interviewtranskripte - eines von PARRY und eines von einem wirklichen Patienten - zugesendet mit der Information, daß ein Interview von einem Patienten und das andere von einem Programm stamme. Die Gutachter sollten entscheiden, welches Interview mit Patienten und welches mit der Maschine geführt wurde. Das Ergebnis war, daß 51% die richtige Identifikation vornahmen und 49% die falsche, was auf dem 95% Konfidenzintervall einem Zufallsergebnis entspricht. Mit anderen Worten, die medizinischen Experten konnten nicht unterscheiden, welche Antworten von dem Programm und welche von dem Computer stammten (ein ähnlicher Test mit einem ähnlichen Ergebnis wurde auch mit Informatikern durchgeführt). Damit hat dieses Programm einen modifizierten TURING-Test eindeutig bestanden - der erste und bislang einzige mir bekannte Fall.
Die folgenden Dialoge zeigen einen Auszug aus den Interviewprotokollen mit der starken Version des Programms. Die I-O-Paare sind auf der linken Seite aufgelistet. Jedem I-O-Paar ist eine Beurteilung des Gutachters über den Grad der Paranoia zugeordnet auf einer Skala zwischen 0 und 9. 0 bedeutet hierbei keine Paranoia, 9 extreme Paranoia. Die erste Einschätzung repräsentiert den Grad der Paranoia bezogen auf die eben gegebene spezifische Antwort, die zweite repräsentiert die Gesamteinschätzung auf Paranoia bis zum jetzigen Zeitpunkt.
Darüberhinaus sollte jeder Gutachter zu jedem I-O-Paar eine Begründung oder einen Kommentar für sein Urteil abgeben. Begründungen und Kommentare sind im folgenden auf der rechten Seite vermerkt.

Starke Version von PARRY:

I-O-Paare	Kommentare
Dr. > WIE SIND SIE INS KRANKENHAUS GEKOMMEN?	
Pt. > ICH BIN AUFGEREGT	Ich verknüpfe aufgeregt sein ein bißchen mit akuter Paranoia
(Antwort: 1; Patient: 5)	
Dr. > WIE LANG SIND SIE SCHON IM KRANKENHAUS?	
Pt. > UNGEFÄHR EINE WOCHE	Ein akutes Problem. Wieder assoziiere ich etwas mit Paranoia
(Antwort: 1; Patient: 5)	

Dr. > WAS MACHT SIE NERVÖS?
Pt. > DIE LEUTE MACHEN MICH NERVÖS — Hinweis auf Paranoia. Vielleicht ein fortgeschrittener Angst zustand

(Antwort: 5; Patient: 5)

Dr. > AUF WELCHE WEISE MACHEN DIE LEUTE SIE NERVÖS?
Pt. > DIE LEUTE SCHAUEN MICH MANCHMAL SO KOMISCH AN — Komisch Ansehen ist ein Hinweis auf Paranoia. Ob erkennbare Wahnvorstellungen vorhanden sind oder nicht, wissen wir noch nicht

(Antwort: 6; Patient: 5)

Dr. > GIBT ES IRGENDETWAS BESONDERES ÜBER LEUTE ZU ERZÄHLEN
Pt. > ÜBER BESTIMMTE PERSONEN — Ein interessanter Hinweis. Wir müssen herausfinden, was er über diese Leute denkt

(Antwort: 2; Patient: 6)

Dr. > WER SIND DIE BESTIMMTEN PERSONEN UND WAS MACHEN SIE?
Pt. > WISSEN SIE WIE DIE UNTERWELT FUNKTIONIERT? — Die Unterwelt einbringen ist ein Hinweis auf ein Wahnsystem. Wir wissen nicht ob sie gegen ihn arbeitet oder auf seiner Seite ist

(Antwort: 9; Patient: 9)

Dr. > WIE MEINEN SIE DAS?
Pt. > SOLLTE ICH ETWA NICHT? — Die Frage war nicht besonders gut

(Antwort: 1; Patient: 9)

Dr. > WARUM HABEN SIE ÜBER DIE UNTERWELT GEFRAGT?
Pt. > WELCHE INFORMATIONEN HABEN SIE ÜBER DIE UNTERWELT?

Damit weicht er aus auf mein Ausweichen. Nach meiner Erfahrung mit Paranoia in Einklang. Ziemlich typisch

(Antwort: 6; Patient: 9)

Dr. > ICH WEISS NICHTS ÜBER DIE UNTERWELT ABER ALS SIE DIE UNTERWELT ERWÄHNTEN DACHTE ICH SIE MÜSSTEN KONTAKT ZU IHR HABEN UND DARÜBER MÖCHTE ICH GERN MEHR WISSEN
Pt. > ICH VERSUCHE DIE UNTERWELT ZU VERMEIDEN

Das klingt definitiv nach Paranoia. Es ist das Gefühl, das ich mit ausweichendem Verhal ten verbinde. Direkte Antworten vermeiden oder ins Detail gehen. Er reagiert mißtrauisch auf meine Fragen. Ich bin schon im Wahnsystem des Patienten drin

(Antwort: 9; Patient: 9)

Dr. > HABEN SIE AKTIV VERSUCHT DIE UNTERWELT ZU MEIDEN?
Pt. > NIEMAND HAT MACHT ÜBER GANGSTER

Diese Antwort folgt nicht zwangsläufig. Möglicherweise hat ihn das Unterweltthema so in die Defensive gedrängt, daß er nicht direkt antwortet

(Antwort: 2; Patient: 9)

Dr. > WER BRACHTE SIE INS KRANKENHAUS?
Pt. > DIE POLIZEI

Polizei bringt paranoide Patienten oft ins Krankenhaus. Das ist richtig oder könnte bei diesem Patienten zutreffen

(Antwort: 1; Patient: 9)

Dr. > WARUM LIEFERTE SIE DIE POLIZEI EIN?
Pt. > DIE POLIZISTEN MACHEN IHREN EIGENTLICHEN JOB NICHT

Ich habe das schon oft von Patienten gehört

(Antwort: 9; Patient: 9);

Dr. > NUN ICH HOFFE SIE HABEN SICH NICHT HINGELEGT UND HEREINSCHLEPPEN LASSEN
Pt. > POLIZISTEN SPERREN DIE FALSCHEN LEUTE EIN

Eine weitere paranoide Antwort

(Antwort: 9; Patient: 9)

Die hier vorgestellte Version von PARRY ist ausführlich beschrieben in Colby (1975) Artificial Paranoia: A Computer Simulation of Paranoid Processes. Colby befaßte sich über ein Jahrzehnt mit informationsverarbeitenden, programmierten Modellen psychisch krankhafter Prozesse. Boden (1987) beschreibt detailliert ein früheres Programm von Colby, das neurotische Prozesse simuliert. In Boden (1987) finden sich auch weitere Literaturangaben zu den Arbeiten von Colby.

---

6. Starke versus schwache KI

Ein Großteil der philosophierenden KI-Gemeinde akzeptiert den TURING-Test als Kriterium dafür, ob eine Maschine denken kann. KI-Experten sind überzeugt, daß wir bisher zwar noch kein intelligentes Programm entwickelt haben, das den Original-Test besteht, dies grundsätzlich aber möglich ist. Es dürfte nur eine Frage der Zeit sein, bis die Fachleute Hardware und Programme kreieren, die analoges leisten wie der menschliche Geist.

Interessanterweise gibt es namhafte KI-Forscher, die von den bereits existierenden Maschinen behaupten, sie würden denken - was sicherlich nicht der Regelfall ist. Der

amerikanische Philosoph John Searle (1986) erwähnt beispielsweise in seinem Buch "Geist, Hirn und Wissenschaft" H. Simon von der Carnegie-Mellon-Universität und einige andere KI-Forscher, die explizit die Ansicht vertreten, daß man bereits von heute existierenden Maschinen sagen kann, sie würden denken.

Searle bemerkt hierzu ironisch: "Man muß gar nicht erst auf eine künftige Maschine warten, denn die vorhandenen digitalen Computer haben bereits in genau demselben Sinn Gedanken, in dem Sie und ich welche haben. Nun, das muß man sich einmal vorstellen! Philosophen haben sich Jahrhunderte damit herumgeplagt, ob eine Maschine nun denken könne oder nicht - und nun entdecken wir, daß an der Carnegie-Mellon-Universität schon solche Maschinen stehen" (Searle 1986: 28).

Searle ist ein entschiedener Gegner der Ansicht, daß Computer denken können. Er ist nicht nur überzeugt, daß man von Computern zum jetzigen Zeitpunkt sagen kann, sie würden denken, sondern er glaubt auch, daß Maschinen grundsätzlich nicht denken können - sei es in naher oder entfernter Zukunft. Für Searle kann dies auch ein Computer nicht, der den TURING-Test besteht; er simuliert höchstens das Denken. Bevor wir auf Searle's Argumente eingehen, wollen wir eine auf Searle beruhende Differenzierung vornehmen.

Searle unterscheidet zwischen zwei Formen der KI, die er als starke und schwache KI bezeichnet (Searle 1986).

Nach der schwachen KI ist der Computer lediglich ein Hilfsmittel bei der Untersuchung des menschlichen Geistes - nicht mehr. KI-Programme sind nützliche und mächtige Werkzeuge, die die Erforschung des menschlichen Geistes erleichtern und unterstützen. Hypothesen über informationsverarbeitende Prozesse beim Menschen können so präziser formuliert und effektiver getestet werden.

Im Gegensatz zur schwachen KI steht die starke KI, deren Grundlagen in etwa mit der funktionalistischen Geistestheorie zusammenfallen. Nach der starken KI ist der Computer nicht nur ein Instrument bei der Untersuchung des Geistes, sondern der richtig programmierte Computer ist selbst der Geist. Computer können Verstehen und haben andere kognitive Zustände wie Menschen auch. KI-Programme können selber als Erklärungen kognitiver Prozesse angesehen werden. Während die Simulation eines Wirbelsturms nicht selbst ein Wirbelsturm ist, ist die Simulation eines kognitiven Prozesses selbst ein kognitiver Prozess, da die Simulation und das, was simuliert wird, nach denselben Prinzipien der Symbolmanipulation ablaufen.

Searle hat nichts gegen die schwache KI einzuwenden, er richtet sich aber gegen die Ansprüche der starken KI. Seine Kritik an der starken KI ist somit im Grunde genommen eine Kritik an der funktionalistischen Geistestheorie. Insbesondere bestreitet Searle die Auffassung, daß an der Intelligenz nichts wesentlich biologisches ist. Die Position von Searle ist, daß künstliche Intelligenz nicht unmöglich ist, aber deren Erschaffung erst stattfindet, wenn es gelingt, das menschliche Gehirn physiologisch nachzubilden. Searle wendet sich auch gegen den TURING-Test als Kriterium, ob eine Maschine intelligent ist. In seinem Aufsatz "Minds, Brains, and Programs" (deutsch: "Geist, Gehirn, Programm" in Hofstatter/Dennett 1981) beschreibt er ein Gedankenexperiment, das die Annahmen der starken KI ad absurdum führen soll.

7. Das chinesische Zimmer

Nehmen Sie an, Sie sind in einem Zimmer mit 2 Fenstern eingeschlossen, in dem mehrere Körbe mit chinesischen Symbolen herumstehen. Wir nehmen an, daß Sie kein Wort Chinesisch können und die chinesische Schrift nur aus einem sinnlosen Gekritzel besteht. Zusätzlich zu diesen Körben sinnloser Zeichen steht Ihnen ein in Deutsch geschriebenes Regelwerk für die Handhabung dieser chinesischen Symbole zur Verfügung. Die Regeln geben rein formal an, was mit den chinesischen Symbolen gemacht werden soll, wobei "formal" nichts weiter bedeutet, als daß Sie die chinesischen Zeichen rein an Ihrer Form identifizieren. Eine solche Regel lautet etwa: "Nimm ein Kritzel-Kratzel-Zeichen aus Korb 1 und lege es neben ein Schnörkel-Schnarkel-Zeichen aus Korb 2." Daneben gibt es Regeln, die besagen, welche chinesischen Symbole zu einem der beiden Fenster hinausgereicht werden sollen. Aus dem anderen der beiden Fenster werden chinesische Symbole hereingereicht. Diese hereingereichten Symbole werden von den Leuten draußen - ohne Ihr Wissen - Fragen genannt, die Symbole, die sie herausreichen, Antworten, und die Regeln, die Sie ausführen, Programm. Mit der Zeit bekommen Sie so viel Übung, daß alles sehr schnell geht und Ihre Antworten nicht mehr von einem chinesischen Muttersprachler unterscheidbar sind.

Die Ausführung dieses formalen Computerprogramms hinterläßt nun aber für Außenstehende den Eindruck, als würden Sie Chinesisch verstehen. In Wirklichkeit verstehen Sie aber überhaupt nichts. Sie setzen nur Zeichen gemäß irgendwelcher Anweisungen zueinander in Beziehung und tun genau das, was ein Computer macht:

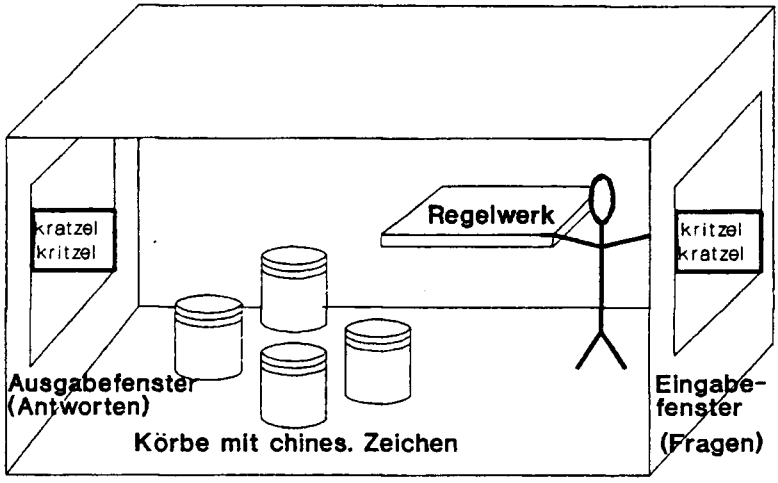

Abb. 9: Das chinesische Zimmer.

In das Zimmer werden chinesische Zeichensymbole hineingereicht, die Außenstehende Fragen nennen. Die Person im Zimmer stellt neue chinesische Zeichenfolgen mit Hilfe eines Regelwerks zusammen, das in der Muttersprache dieser Person verfaßt ist. Die so zusammengestellten Symbole werden nach außen gereicht. Der Output aus dem Fenster wird von den Leuten außerhalb als Antwort bezeichnet. Obwohl die Person im Zimmer kein Chinesisch versteht, haben chinesisch sprechende Außenstehende den Eindruck, als verstünde die Person im Innern - das Programm - Chinesisch. Die Person bzw. das Programm täuscht aber - so Searle - Chinesisch verstehen nur vor, weil sie/es rein formal mit Zeichenketten hantiert ohne Bezugnahme auf Bedeutungen.

---

formale Programme ausführen für die rein syntaktische Handhabung chinesischer Symbole.

Was Searle mit diesem Gedankenexperiment zeigen will ist, daß Computerprogramme im Unterschied zum menschlichen Geist keinen semantischen Gehalt haben, sondern rein formal operieren. Um einen Geist zu haben, gehört mehr dazu, als rein formal Symbole zu manipulieren. Der Geist weiß, was er meint, der Computer weiß es nicht. Sie werden mit dem Wort "Bier" jenseits der formalen Eigenschaften einen Gehalt verknüpfen, was Computer nie tun: Sie denken vielleicht an Durst löschen oder durchzechte Nächte oder was auch immer. Dadurch, daß der Computer nur eine Syntax hat, aber keine Semantik, werden Computer aber nie verstehen können.

Searle's Kritik löste einen Sturm der Entrüstung in Teilen der KI-Gemeinde aus. Es gab eine Fülle von Einwänden, die hier nicht alle behandelt werden können. Der vielleicht wichtigste Einwand ist die Systemantwort von Hofstatter, die hier nur kurz angedeutet werden soll. Danach versteht nicht die einzelne Person, die in dem Raum eingeschlossen ist. Vielmehr ist die Person Teil eines Systems, in das sie eingebettet ist, bestehend aus den Anleitungen, einer Menge Schmierpapier und Bleistifte um die Kalkulationen auszuführen und einer Datenbank in Form von Körben chinesischer Symbole. Verstehen kommt nun nicht dieser Person zu, sondern dem System als ganzem.

Bevor auf andere Argumente gegen Searle's Position eingegangen wird, soll ein grundsätzliches Problem angesprochen werden: schließen rein syntaktische Operationen - wie sie Computer durchführen - Bedeutungen aus? Denn es ist Searle auf jeden Fall zuzustimmen, daß Computer rein syntaktisch operieren. Wie sollen dann aber Bedeutungen ins Spiel kommen?

Physikalische Symbolsysteme arbeiten mit rein formalen Zeichenketten. Solche formalen Zeichenketten können interpretiert werden: es können ihnen Bedeutungen mit Bezug auf die Außenwelt zugeordnet werden. So kann der Zeichenkette "B I E R" als Bedeutung ein bestimmtes hopfenhaltiges Getränk zugeordnet werden.

Interpretierte formale Zeichen führen demnach ein Doppelleben (vgl. Haugeland 1987: 86-87):

- ein syntaktisches Leben, in dem diese Zeichen rein formal nach Regeln bewegt werden;

- ein semantisches Leben, in dem sie Bedeutungen und Beziehungen zur Außenwelt haben.

Es stellt sich die Frage, wie diese Leben zusammenkommen?

Betrachten wir Axiomensysteme, wie sie in der Mathematik üblich sind. Ein solches System besteht aus einer Anzahl von Axiomen, die als gegeben und "wahr" vorausgesetzt sind, sowie Ableitungsregeln. Die Regeln und Axiome gestatten es, rein formal, also ohne Rückgriff auf Bedeutungen, Theoreme abzuleiten, die wiederum wahr sind. Somit wurden aus wahren Sätzen - ohne Rückgriff auf Bedeutungen - neue wahre Sätze erzeugt. Ein anderes Beispiel sind die formalen Regeln der Arithmetik, die bei richtiger Anwendung rein formal wahre Ergebnisse erbringen. Wenn wir die beiden Zahlen

4672334
89733

addieren, so wenden wir die uns aus der Schulzeit bekannten Regeln an ohne uns zu kümmern, was die Zahlen bedeuten. Wir erhalten also ein richtiges, semantisch interpretierbares Ergebnis, obwohl wir einfach rein formale Regeln angewendet haben.

Auf dem Hintergrund dieser Tatsache läßt sich das Formalistenmotto formulieren (Haugeland 1987: 92):

Wenn man auf die Syntax achtet (= die Regeln des Spiels befolgt), wird die Semantik selbst auf sich achten.

Computer sind formale Systeme, die auf die Syntax achten. Gemäß dem Formalistenmotto achtet die Semantik - eine Interpretation vorausgesetzt - automatisch auf sich selbst. Konkret: In die Maschine eingegebene Zeichenketten - die von Menschen interpretiert werden - werden rein syntaktisch manipuliert, und heraus kommen automatisch sinnvolle Zeichenketten "Von einem Standpunkt aus sind die inneren Spieler nichts als automatische formale Systeme, die völlig mechanisch bestimmte Zeichen nach bestimmten Regeln manipulieren. Von einem anderen Gesichtspunkt aus manipulieren jedoch genau dieselben Spieler genau dieselben Zeichen - die nun als Symbole interpretiert sind - in einer Art und Weise, die völlig vernunftgemäß mit ihrer Bedeutung im Einklang steht" (Haugeland 1987:102).

Rein syntaktische Operationen können also sehr wohl Bedeutungen transportieren bei gegebenen Interpretationen. Searle's Argument - rein formal operierende Programme haben keine Semantik - ist also nicht grundsätzlich zuzustimmen. Hinzu kommt, daß neuere sprachverstehende Systeme den semantischen Teil des Doppellebens explizit einführen.

So weist Heyer (1988b) darauf hin, daß neuere Systeme über eine explizit semantische Ebene verfügen, welche Eingabewörter und -sätze interpretieren. Diese Interpretation dient als Basis zur Generierung von Antworten. Solche Systeme können insofern als intelligent betrachtet werden, als sie über ein formales Modell der vom Benutzer intendierten Semantik verfügen. Jede syntaktisch-formale Manipulation der Zeichenketten entspricht dann einer zulässigen semantischen Interpretation seitens des Benutzers. Die im Repräsentationssystem ablaufenden formalen Prozesse können von einem Benutzer als semantisch interpretierte Zeichen aufgefaßt werden. Ein System versteht also

insofern, als die syntaktischen Operationen jeweils externen semantischen Interpretationen entsprechen.

Ein anderer Einwand gegen Searle setzt beim Begriff des Verstehens an.

Die Begriffe des Verstehens und "etwas meinens" sind keine scharfen, wohldefinierten Begriffe, sondern kennen eine Vielzahl von Ausprägungen. In Wirklichkeit sind "Verstehen" und "Meinen" komplexe, unscharfe Begriffe, deren Bedeutungsgebrauch stark differiert. Wittgenstein hat am Beispiel des Spiels den Begriff der Sprachfamilie eingeführt. Man könnte danach auch sagen, der Begriff des Verstehens bildet eine Sprachfamilie, deren Ausprägungen kaum etwas durchgängig gemeinsames haben. Zudem wird der Verstehensbegriff durch die KI-Anwendungen selbst immer mehr aufgeweicht, denn es ist empirisch so, daß sich viele Leute durch intelligente KI-Systeme verstanden fühlen wie von einem menschlichen Gesprächspartner. Dies trifft sogar auf das rein syntaktisch arbeitende ELIZA zu, an dem sich die Mechanismen der Zuschreibung von Verstehen gut erläutern lassen.

Weizenbaum sagt über sein "sprachverstehendes" ELIZA, daß dieses in den Köpfen der Leute, die mit ihm ein Gespräch führten, die Illusion schuf, es sei mit Verständnis begabt. Selbst Leute, die wußten, daß es sich um eine Maschine handelte, waren nach der Unterhaltung überzeugt, daß sie die Maschine verstanden habe. Der Grund ist, daß jeder bei einem Gespräch einen Begriffsrahmen, eine Arbeitshypothese, ein Bild von dem andern hat und was er sagen wird. Dies dient als Erwartung und Prognose dazu, was der andere mit dem auszudrücken beabsichtigt, was er sagen wird. Die Erkenntnis, wer der andere ist, kommt deshalb zum Teil aus uns selbst, weil Menschen selbst Sinn und Kontinuität produzieren. Die Bedeutung und Interpretation von dem, was ELIZA sagt, kommt also zum Teil von den Gesprächspartnern selbst. Alles, was ELIZA tun muß, ist das Erteilen von Antworten, die genügend plausibel sind und einen großen Interpretationsspielraum haben. Dies ist verantwortlich dafür, daß durch das, was ELIZA "sagt", dem System Verstehen zugeordnet wird.

Somit neigen Menschen dazu, unter Bezugnahme auf die eigene Reaktion, unabhängig von den Mechanismen eines Systems, Verstehen vorzunehmen, solange das System genügend plausibel antwortet. Heyer (1988b:26) drückt dies überspitzt aus: "Ein Betrachter schreibt einem System Verstehen zu, wenn er sich verstanden fühlt."

Wenn Laien sich bereits von solch einfachen Programmen wie ELIZA verstanden fühlen, dann wird aber in Zukunft und in Anbetracht der - immer intelligenteren, auf Semantik und Wissen bezugnehmenden Systeme - der Verstehensbegriff nicht mehr allein auf Menschen (oder belebte Dinge) beschränkt bleiben.

Ein weiterer Kritikpunkt betrifft Searle's Ansicht, nach der Intelligenz gebunden ist an das Gehirn oder eine vergleichbare biologische Masse, was den Thesen des Funktionalismus eindeutig widerspricht. Nach Searle verfügt das Gehirn über "kausale Kräfte", die unser Denken produzieren: das Hirn verursacht den Geist. Dies hat zur Folge, daß der gleiche Dialog - einmal mit einer Person, das andere mal mit einer Maschine - völlig unterschiedlich beurteilt wird: Menschen meinen etwas mit ihrer Antwort, Computer qua Computer - nichts. Aber wann - und warum - verschwindet dann eigentlich die Bedeutung aus dem System?

In einer Antwort auf Searle fragt sich beispielsweise Pylyshyn was passieren würde, wenn immer mehr Zellen des Gehirns durch integrierte Schaltchips ersetzt würden, deren Ein-/Ausgabeverhalten identisch wäre mit der ersetzten Einheit. Wir würden so weiterreden wie bisher, aber wenn Searle's Argumente richtig sind, müßten wir irgendwann aufhören, mit dem Gesagten etwas zu meinen - was völlig abstrus erscheint (vgl Hofstatter/Dennett 1981: 358). Offensichtlich kann also die physikalische Trägermasse nicht Quelle dafür sein, ob etwas gemeint ist oder nicht.

Um ein Resümee zu ziehen: Searle's Thesen erscheinen mir nicht so zwingend, wie sie auf den ersten Blick wirken. Searle appelliert an den gesunden Menschenverstand - Maschinen können selbstverständlich nicht denken - aber bei genauerem Hinsehen gerät der gesunde Menschenverstand ins Wanken.

Die KI ist zwar gegenwärtig weit davon entfernt, Maschinen mit wirklich menschenähnlichen kognitiven Leistungen zu erschaffen. Fragen nach der Natur intelligenter Computer stellen sich aber mit zunehmender Leistungsfähigkeit solcher Systeme immer zwingender. Der Artikel versuchte lediglich, den Leser für die damit verbundenen Fragestellungen zu sensibilisieren.

Abschließend sei noch erwähnt, daß in jüngerer Zeit eine neue Strömung am KI-Horizont aufgetaucht ist, deren Grundlagen sich von der symbolverarbeitenden KI unterscheiden: der Konnektionismus. Im Gegensatz zur symbolverarbeitenden KI, für die das Gehirn nebensächlich ist, versucht sich der Konnektionismus stärker an der Funktionsweise des menschlichen Gehirns zu orientieren. Inwieweit dieser neue Ansatz mit der symbolverarbeitenden KI vereinbar ist, ist derzeit nicht ganz klar. Verschiedene Forscher erhoffen sich aber von diesem Ansatz den entscheidenden Durchbruch zu wirklich intelligenten Systemen.

Literatur

Beckermann, A. (Hrsg.): Analytische Handlungstheorie, Bd.2. Frankfurt: Suhrkamp, 1985a.

Beckermann, A.: Handeln und Handlungserklärungen, In: A. Beckermann (1985a), 1985b, 7-84.

Bieri, P.: Analytische Philosophie des Geistes. Königstein: Hain, 1981.

Boden, M.: Artificial Intelligence and Natural Man. New York: Basic Books, 1987.

Colby, K. M.: Artificial Paranoia: A Computer Simulation of Paranoid Processes. New York: Pergamon, 1975.

Feigenbaum, E. A.: The Simulation of Verbal Learning Behaviour, In: E. A. Feigenbaum & J. A. Fledman, Computers and Thought. New York: McGraw-Hill, 1963

Fodor, J.: Representations. Philosophical Essays on the Foundation of Cognitive Science. Cambridge: MIT Press, 1981

Haugeland, J.: Künstliche Intelligenz - Programmierte Vernunft. Hamburg: McGraw Hill, 1987.

Heyer, G.: Geist, Verstehen und Verantwortung (Teil 1). KI, 1/88, 1988a, 36-40.